Deep learning models for identifying crop and field attributes from near surface cameras

Shawn D. Taylor[1,2*] and Dawn M. Browning[1]

[1]US Department of Agriculture, Agricultural Research Service, Jornada Experimental Range, New Mexico State University, Las Cruces, New Mexico, 88003, USA

[2]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee, 37830, USA

[*]Corresponding author. Email: shawn.taylor@usda.gov

Keywords: crop phenology; PhenoCam; VGG16; hidden markov model; time series; LTAR

Abstract

Near surface cameras, such as those in the PhenoCam network, are a common source of ground truth data in modelling and remote sensing studies. Despite having locations across numerous agricultural sites, few studies have used near surface cameras to track the unique phenology of croplands. Due to management activities, crops do not have a natural vegetation cycle which many phenological extraction methods are based on. For example, a field may experience abrupt changes due to harvesting and tillage throughout the year. A single camera can also record several different plants due to crop rotations, fallow fields, and cover crops. Current methods to estimate phenology metrics from image time series compress all image information into a relative greenness metric, which discards a large amount of contextual information. This can include the type of crop present, whether snow or water is present on the field, the crop phenology, or whether a field lacking green plants consists of bare soil, fully senesced plants, or plant residue. Here we developed a modelling workflow to create a daily time series of crop type and phenology, while also accounting for other factors such as obstructed images and snow covered fields. We used a mainstream deep learning image classification model, VGG16. Deep learning classification models do not have a temporal component, so to account for temporal correlation among images our workflow incorporates a hidden markov model in the post-processing. The initial image classification model had out of sample F1 scores of 0.83-0.85, which improved to 0.86-0.91 after all post-processing steps. The resulting time series show the progression of crops from emergence to harvest, and can serve as a daily, local scale dataset of field states and phenological stages for agricultural research.

1 Introduction

The timing of planting, emergence, maturity, and harvest of crops affects the yield and long-term sustainability of croplands, thus tracking crop phenology has numerous interested parties from local to national levels. Remotely sensed data is an important data source for tracking crop phenology, and other attributes, from the field to global scales (*Weiss et al.*, 2020). Estimating crop phenology from remotely sensed data is difficult since crops do not follow the same patterns as natural vegetation. Harvest may happen when crops are still green, and multiple harvests in a single year will result in several "peaks" in greenness. Crop rotations result in different crop types year to year, which affects the relative greenness and derived phenology metrics (*Gao et al.*, 2017). A primary limitation for improving satellite remote sensing based crop phenology models is a lack of widespread ground truth data (*Gao and Zhang*, 2021). Near-surface cameras used in the PhenoCam network offer a novel solution to this need for local-scale crop management information. Images from near-surface cameras can document the date of emergence, maturity, harvest, and tillage at the field scale with a daily temporal resolution (*Hufkens et al.*, 2019).

Automatically extracting cropland attributes and phenological information from PhenoCam imagery is difficult. The primary method to estimate phenological metrics with PhenoCam data uses the direction and amplitude of a greenness metric (the green chromatic coordinate, *Gcc*) of regions of interest in the camera field of view (*Richardson et al.*, 2018b; *Seyednasrollah et al.*, 2019). These metrics are well correlated with crop emergence and maturity, but cannot be used to directly identify other attributes such as crop type, flowering, or the presence of crop residue (*Browning et al.*, 2021). Deep learning models provide a straightforward method for identifying information in images, and can potentially identify phenological states directly as opposed to inferring them from relative greenness in the images. Studies have successfully used deep learning image classification models to identify and count animals (*Weinstein*, 2018; *Norouzzadeh et al.*, 2018), classify animal movement (*Conway et al.*, 2021), and identify the phenological stage (*Correia et al.*, 2020), species (*Jones*, 2020), or stressors (*Ghosal et al.*, 2018) of individual plants. Deep learning has been used previously with PhenoCams to identify images with snow cover with up to 98% accuracy (*Kosmala et al.*, 2016).

In agriculture, deep learning classification of near surface images, either from fixed or handheld cameras, has primarily been used for weed and crop disease detection (*Benos et al.*, 2021). Few studies have used deep learning for the classification of crop and field attributes (e.g. *Yalcin* (2017); *Han et al.* (2021)) and to our knowledge no study has applied deep learning methods at cropland sites in the PhenoCam image archive (*Richardson*, 2019).

Here we use images from 55 agricultural cameras in the PhenoCam network to build a classification model for identifying cropland phenological states. A variety of crops are used to generalize the states into 21 classes across 3 mutually exclusive categories, ranging from emergence to harvest. We also included classes for crop type and factors such as flooded or snow covered fields. Deep learning models designed for image classification do not have a temporal component, so we use a hidden markov model in the classification post-processing to account for the temporal correlation of daily camera time series. Results show the feasibility of a daily, local scale dataset of field states

and phenological stages for agricultural research.

2 Methods

2.1 Data

We used PhenoCam images from agricultural sites to train an image classification model (Figure S1, Table S1). To obtain a representative sample of images across all potential crops and crop stages we used seasonal transition dates provided by the PhenoCam Network. Based on the transition date direction (either rising or falling) and threshold (10%, 25%, and 50%) we partitioned each calendar year into distinct periods of senesced, growth, peak, and senescing (*Richardson et al.*, 2018b). We chose 50 random days from each site, year, and period, for a total of 8,270 images. We annotated each image by hand using the imageant software into the 21 classes described below (*Barve et al.*, 2020).

Initial image classifications were organized into 21 classes across three categories of Dominant Cover, Crop Type, and Crop Status (Table 1). The categories are each mutually exclusive such that any single image can be independently classified into a single class within each category. This allows finer grained classification given an array of Crop Types, and more flexibility in classifying crop phenological stages. For example, it would be informative for remote sensing models to know the exact date of crop emergence, but also that on the specified date and for several days to weeks after the feld is still predominantly bare soil. The first category, Dominant Cover, is the predominant class within the field of view. The Crop Type category represents the four predominant crops in the dataset (corn, wheat/barley, soybean, and alfalfa). Wheat and barley are combined into a single category as they are difficult to discern in images. The unknown Crop Type class is used during emergence when an exact identification is impossible. The other Crop Type class represents all other crops, including fallow fields, besides the four predominant ones. The stages of the Crop Status category are loosely based on BBCH descriptions (*Meier*, 1997), but generalized to be applicable across a variety of crops and what is discernible. The Flowers stage is used for identifying tassels on corn or seed head on wheat or barely since no other reproductive structures were visible in the images. Thus, crop types other than corn and wheat/barley do not have any images annotated as Flowers.

After annotation we excluded some images based on low prevalence of some category combinations. For example, only 8 images had the combined combination of Soil, Unknown Plant, and Senescing for the Dominant Cover, Crop Type, and Crop status categories, respectively. When a unique combination of the three categories had less than 40 total images, all images representing that combination were excluded from the model fitting. This resulted in 255 annotated images, from the original 8,270, being excluded. A total of 8,015 annotated images were available for model fitting.

We used mid-day images in the annotation stage and leveraged more of the PhenoCam archive to increase sample size for model fitting. The 8,015 images that we annotated represent the mid-day image for a single date, though phenocams record images up to every 30 minutes. For each

Category	Class	Used in final product	Description
Dominant Cover	Blurry	No	Image blurry, out of focus, or otherwise obscured.
	Vegetation	Yes	Live or senesced vegetation
	Residue	Yes	Post-harvest plant residue
	Bare soil	Yes	
	Snow	Yes	
	Water	Yes	
Crop Type	Blurry	No	Image blurry, out of focus, or otherwise obscured.
	Unknown Plant	Yes	Plants are present but cannot be confidently identified
	Corn	Yes	
	Wheat/Barley	Yes	
	Soybean	Yes	
	Alfalfa	Yes	
	Other	Yes	Any other crop or a fallow field.
	No crop	Yes	No crop present (eg. a plowed field or completely snow covered)
Crop Status	Blurry	No	Image blurry, out of focus, or otherwise obscured.
	Emergence	Yes	First shoots and/or leaves are visible.
	Growth Stage	Yes	Plants have several distinct leaves and/or tillers visible, but no visible tassels,flowers,or fruit.
	Tassels/Flowering	Yes	Plants have distinct tassels, flowers, or fruit.
	Senescing/browning	Yes	10% or more of visible plants are brown/browning.
	Fully senesced	Yes	90% or more of visible plants are fully senesced.
	No crop	Yes	No crop present (eg. a plowed field or completely snow covered)

Table 1: Class descriptions used in the classification model.

annotated image date we also downloaded all images between 0900 and 1500 local time, resulting in an additional 83,469 images. We applied the annotation of the midday image to all images of that date, resulting in 91,484 total images used in the model fitting. This allowed us to increase training image data by a factor of 10 with minimal effort, and include more variation in lighting conditions. While it's possible some of these non-midday images were annotated incorrectly (e.g., a blurry camera becoming cleared, or a field being plowed after midday) these are likely minimal and did not have a large effect on model accuracy (*Norouzzadeh et al.*, 2018).

2.2 Image Classification Model

We used the VGG16 model in the Keras python package to classify images into the 21 classes (Table 1)(*Simonyan and Zisserman*, 2014; *Chollet*, 2018). The model allows us to specify the hierarchical structure of the three categories, such that the predicted class probabilities for any image sum to one within each category. We held out 20% of the images as a validation set. The validation set included all images from three cameras: arsmorris2, mandani2, and cafboydnorthltar01 totaling 10,172 images. It also included 8,124 randomly selected images from the remaining locations to obtain the full 20%. This resulted in a validation sample size of 18,296 and a training sample size of 73,188. The VGG16 classification model was trained fully, as opposed to using transfer learning (*Norouzzadeh et al.*, 2018). We experimented with transfer learning, where a pre-trained model is fine-tuned using our own data, but found that training the model fully had better results.

We resampled the 73,188 training images to 100,000 using weights proportional to the unique combinations among the three categories. For example, there were 4,407 images annotated as Vegetation, Corn, and Flowers for the three categories, but only 1,074 images annotated with Vegetation, Wheat, and Flowers. The images in the former class were given a lower weight in the resampling to reach 100,000 total training images. This allows for even sample sizes among classes and protects against the model being biased toward common classes. During model fitting the images are shuffled and transformed using random shifts and rotations such that the exact same image is never seen twice, which protects against over-fitting. We trained the model with an image resolution of 224x224 pixels using the Adams optimizer with a learning rate of 0.01 for 15 epochs, and an additional 5 epochs with a learning rate of 0.001.

2.3 Post-processing

After fitting, the VGG16 model was used to classify 55k midday images from all agricultural PhenoCam sites , totalling approximately 170 site-years (*Milliman et al.*, 2019). These classified images were then put through a post-processing routine to produce a final classification for each day (Table 2). First, for all dates marked as Snow in the Dominant Cover category, the Crop Type and Crop Status predictions were removed and gap-filled using linear interpolation from surrounding non-snow dates as long as the gap was 60 days or less. The reasoning behind this is during the constant snow cover of winter the crop, if any, likely remains unchanged. Next, any image marked as blurry was removed and the associated image date marked as missing across all three categories (Dominant Cover, Crop Type, Crop Status). Gaps of missing dates, up to 3 days, in the time series

1. Predict probabilities for each of the 55k daily images.
2. For snow days remove Crop Type and Crop Status and gap-fill up to 60 days.
3. For blurry images remove all predictions for that date and gap fill up to 3 days.
4. Apply HMM to Dominant Cover and Crop Status categories.
5. Identify each unique crop sequence. A crop sequence is all dates between two non-consecutive "no crop" dates of the Crop Status category.
6. For each crop sequence identify the Crop Type using the highest cumulative probability excluding the unknown class.
7. Mark crop sequences as unknown for sequences 60 days or less, where the most common Crop State was emergence.
8. Mark crop sequences as unknown when the highest cumulative probability within the sequence was No Crop.

Table 2: The post-processing steps performed on the VGG16 model predictions.

for each site were filled with a linear interpolation of the two bounding date probabilities for each of the remaining 18 classes. Probabilities across all dates were normalized across the remaining classes to account for removing the "Blurry" class. After this initial filtering the classification time series from the 55k images were input into an hidden markov, and several additional post-processing steps, to produce the final time series.

Out-of-the box image classification models such as VGG16 have no temporal component. Every image is treated as an observation independent of temporally adjacent images. Thus misclassification of images can lead to noisy time series and improbable transitions between classes. To correct for this we used a hidden markov model (HMM) to reduce the day-to-day variation and remove improbable transitions (*Esmael et al.*, 2012; *Wehmann and Liu*, 2015). HMM's are state-space models which combine a latent "true" state of a process with an observation model. The latent state evolves dynamically where every timestep is a discrete state which depends only on the state of the previous timestep, and where the probability of moving from one state to another is decided by a transition matrix. The observation model is a timeseries of the same length where every observed state depends only on the latent state of the same timestep. Given a latent state, all observed states have a non-zero probability of being observed.

We used two HMMs, one for Dominant Cover and another for Crop Status, each with a daily timestep. Only sequences with at least 60 continuous days, after the gap-filling from Snow and Blurry dates described above, were processed with the HMM. For the observation model within each HMM we used the direct output from the classification step, which for each image date consisted of probabilities of the image belonging to each class in the respective category. The transition matrix describing the probabilities of the hidden state changing states from one day to the next was created manually for each HMM (Table S2-3). The probability of the hidden state staying the same between two dates was set as the highest (0.90 - 0.95) in all cases. In this way the hidden state only changes when there is strong evidence in the observation model, represented by continuous and high observed probabilities of a new state. Within the Dominant Cover category, transitions to other states besides the current one were set to equal values with one exception. Transition from Soil to Residue was

set to 0 probability since residue can only be present after a crop has been harvested. Within Crop Status transitions between states were constrained to be biologically possible. Improbable transitions (e.g., from Emergence to Senescing in a single day) had probabilities of 0. Reproductive structures were not visible on all crops in images, so transitioning from the Growth stage directly to Senescing was allowed. Transitioning from either Senescing or Senesced to Growth was also allowed since this represents dormancy exit in overwintering crops such as winter wheat. Given observation probabilities and the transition matrices the most likely hidden state was predicted using the viterbi algorithm to produce the final time series across the Dominant Cover and Crop Status categories.

For the Crop Type category the HMM methodology is less useful, since no day-to-day transitions between Crop Types are expected. Here we used the HMM output of Crop Status to identify each unique crop sequence, defined as all dates between any two "No Crop" classifications. For each unique crop sequence, we identified the associated Crop Type with the highest total probability within that sequence, and marked the entire sequence as that Crop Type. In this step we considered all Crop Type classes except "Unknown Plant", which is used during the emergence stage when a plant is present but the exact type is unclear. This allows information later in the crop cycle, when Crop Type is more easily classified, to be propagated back to the emergence stage. Finally, we marked the final Crop Type as Unknown for some crop sequences in two instances: 1) when the length of a crop sequence is less than 60 days and the sequence was predominantly in the emergence stage, and 2) when the "No Cop" class was selected as the final Crop Type with the highest probability within a sequence. These two scenarios tended to occur when volunteer plants are growing sparsely on an otherwise bare field.

2.4 Evaluation

We calculated three metrics to evaluate the performance of the image classifier: precision, recall, and the F1 score. All three metrics are based on predictions being classified into four categories of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Precision is the probability that an image is actually class i, given that the model classified it as class i. Recall is the probability that an image will be classified into class i, given that the image is actually class i. The F1 score is the harmonic mean of precision and recall. All three metrics have the range 0-1, where 1 is a perfect classification.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

where TP, FP, FN are the number of true positive, false positive, and false negative classifications, respectively. Using only the mid-day images we calculated all three metrics for the 21 original classes to evaluate the performance of the VGG16 model. For this step the predicted class

for each category on a single date was the one with the maximum probability. We re-calculated all three metrics again after all post-processing steps. The second round of metrics does not include two classes removed during post-processing: the Blurry class across all categories and the Unknown Plant class in the Crop Type category.

Software packages used throughout the analysis include keras (*Chollet*, 2018), tensorflow (*Abadi et al.*, 2016), pandas (*McKinney*, 2010), numpy (*Harris et al.*, 2020), and pomegranate (*Schreiber*, 2018) in the python programming language (*Python Software Foundation*, 2003). In the R language (*R Core Team*, 2017) we used the zoo (*Zeileis and Grothendieck*, 2005), tidyverse (*Wickham et al.*, 2019), and ggplot2 (*Wickham*, 2016) packages. All code for the analysis, as well as the final model predictions, are available in a Zenodo repository (https://doi.org/10.5281/zenodo.5579797).

3 Results

The overall F1 score, a summary statistic which incorporates recall and precision, was 0.90-0.92 for the training data across the three categories of Dominant Cover, Crop Type, and Crop Status (Figure 1). The overall F1 score for validation data, which was not used in the model fitting, was 0.83-0.85 for the three categories. In the Dominant Cover category the vegetation class was the best performing overall with recall and precision of 0.97 and 0.93, respectively. Thus the classification model has a strong ability to discern when the camera field of view is or is not predominantly vegetation. When vegetation is not dominant the classifier is still moderately accurate, though there is confusion between soil and residue classes indicated by their recall scores (0.64-0.68). The precision of soil and residue was 0.61 and 0.82 for validation data, indicating that the classifier leaned toward residue.

Excluding the blurry class, the worst precision for Crop Type was soybeans, with a precision of 0.53 on the validation data. This indicates that if an image was classified as Soybean, then there is a 53% chance it is actually soybean. The recall for Soybean was high, 0.97 with the validation data, indicating that there is a high amount of false positives from non-Soybean images being classified as Soybean. Conversely the Wheat/Barley and Other classes have high precision (0.86 and 0.86, respectively), and low recall (0.60 and 0.65, respectively). This indicates a high amount of false negatives, where images of Wheat/Barley and Other Crop Types are being classified as other Crop Type classes.

The blurry class had low recall and precision across all three categories, with values of 0.29-0.43 and 0.33-0.50 for validation data recall and precision, respectively. Combined with training data recall scores of 1.0, this indicates likely over-fitting of the blurry class in the classification model. We did not attempt to improve this further since the blurry image prevalence was extremely low. Additionally, when images were marked as blurry in the final dataset the final state was interpolated in the HMM post-processing step by accounting for the surrounding images.

Figure 2 shows the classification statistics after post-processing of the image time series, where the HMM was used for Dominant Cover and Crop Status and the Crop Type was set to the highest total probability within any single crop series. The blurry class is not shown here since it was removed in the post-processing routine. The Unknown Plant class for Crop Type is also excluded since in

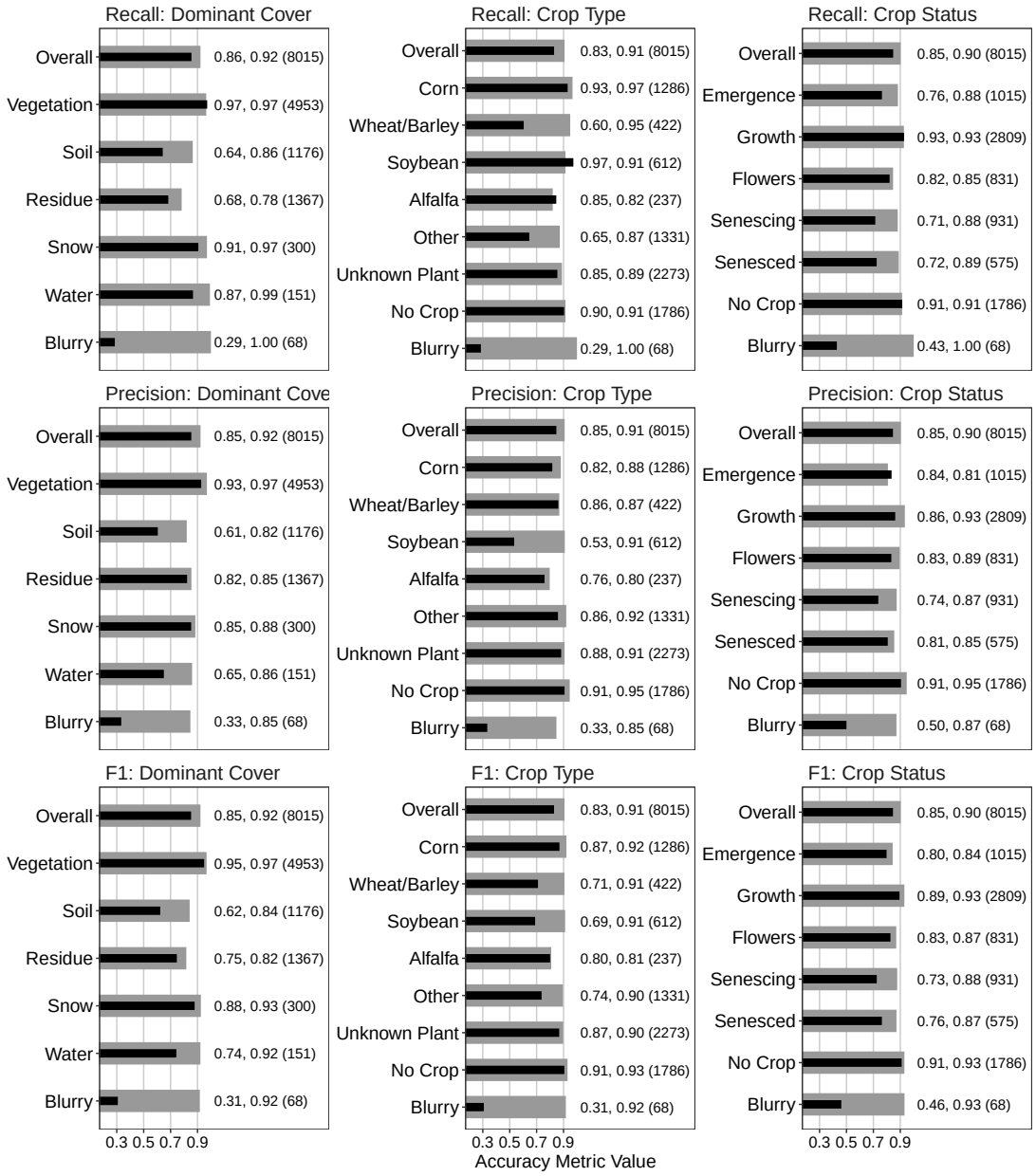

Figure 1: Accuracy metrics for the VGG16 image classifier. The black and grey bars represent the validation and training datasets, respectively. The text indicates the respective metric value for validation, training, and class sample size in parentheses. The training and validation sample sizes are 80% and 20% of the total sample size, respectively. Overall indicates the average metric value for the respective category, weighted by sample size. All three metrics have a range of 0-1 where 1 equals a perfect prediction.

the post-processing the Crop Type category is assigned to the highest probability class seen in each crop sequence, thus performance metrics for the Unknown Plant class would be uninformative.

The validation data performance metrics after the post-processing steps either improved or re-

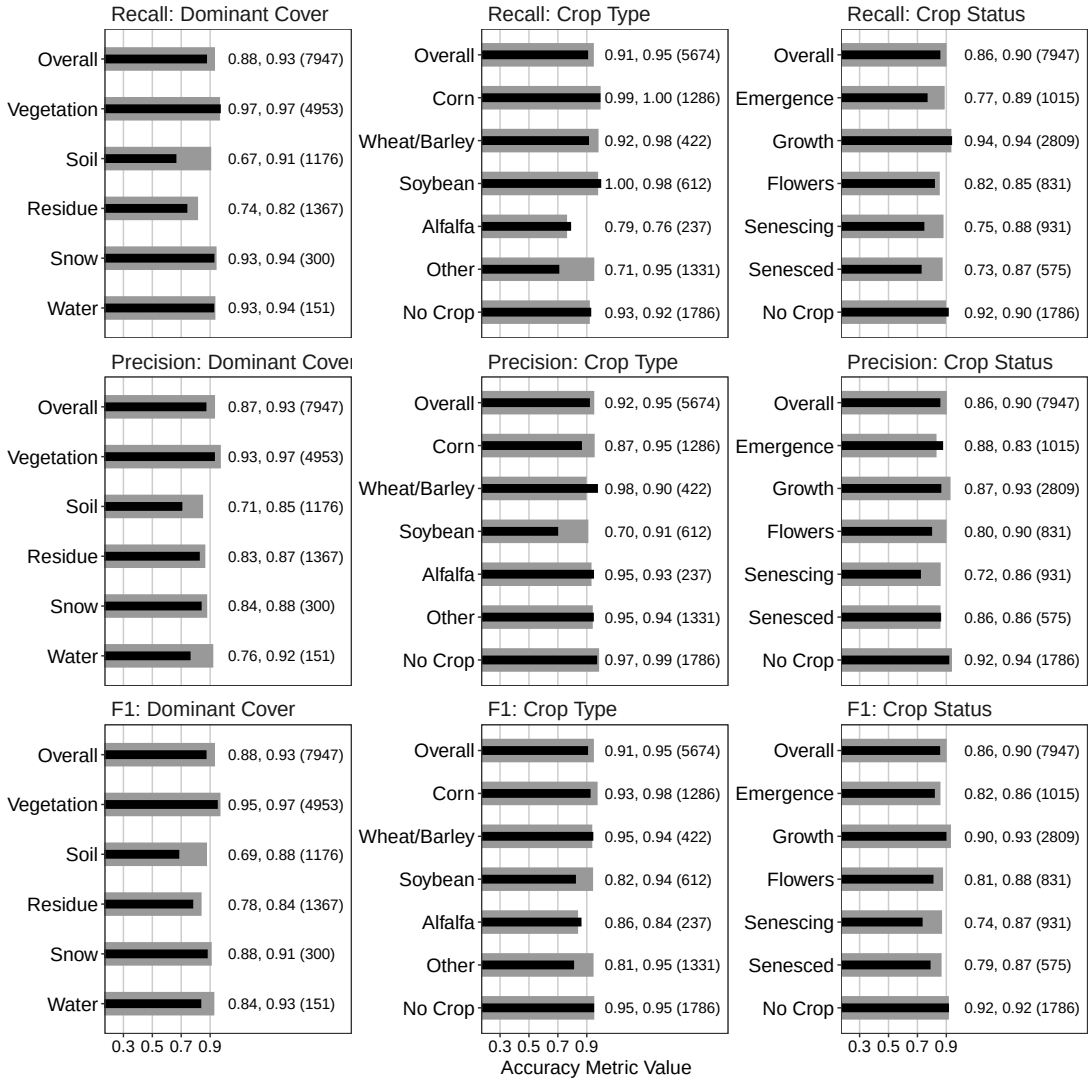

Figure 2: Accuracy metrics for the classifications after post-processing. The black and grey bars represent the validation and training datasets, respectively. The text indicates the respective metric value for validation, training, and class sample size in parentheses. The training and validation sample sizes are 80% and 20% of the total sample size, respectively. Overall indicates the average metric value for the respective category, weighted by sample size. All three metrics have a range of 0-1 where 1 equals a perfect prediction. Differences between this and Figure 1 is the exclusion of the blurry and unknown plant classes, with total sample sizes reflecting this.

mained the same across all classes except four: the Snow class in Dominant Cover, the Alfalfa Crop Type, and Flowers and Senescing Crop status classes. Overall F1 scores using validation data increased from 0.85 to 0.88, 0.83 to 0.91, and 0.85 to 0.86 for Dominant Cover, Crop Type, and Crop Status categories, respectively.

Next we present four examples of the full classification and post-processing results using a single calendar year from four sites. They show the initial output of the classification model, as well as

the capability of the HMM in removing high variation in the original VGG16 model prediction. The original VGG16 predictions are obtained by choosing the class with the maximum probability (MaxP) for each category and day. We compare them with insight gained from the full image time series available on the PhenoCam data portal https://phenocam.sr.unh.edu. For example at the arsmorris2 site in central Minnesota, in the months March through June of 2020, there is uncertainty in whether the Dominant Cover of the field is Residue or Soil (Figure 3A, MaxP). The HMM model resolved it to the Residue class for the three month period. From mid-June thru October there is high certainty that that vegetation is present, reflected in both the initial classifications (MaxP) and resulting HMM. The HMM model resolved the Crop Type as Corn (Figure 3B). During June to October, the Crop Status progresses naturally through the different stages, and uncertainty arises only in October when fully senesced vegetation is confused with plant residue (Figure 3C).

Cropping systems with multiple harvests per season are challenging for remote sensing models. There were multiple harvests for the bouldinalfalfa site in northern California for the year 2018. Here an alfalfa field was persistent for the entire year with several harvests (Figure 4A). During the intervals of regrowth after each harvest the Dominant Cover of the field was classified as Residue with emergence of an Unknown crop type (Figure 4B). Once the plants matured then it was identified consistently as Alfalfa, which in the post-processing was back propagated in time for each crop sequence.

At the site cafcookeastltar01 in eastern Washington in the year 2018 there was a short residual crop of wheat in April (Figure 5A). Since the plants were not allowed to grow into the summer, due to a new crop being planted, they were not positively identified and instead marked as Unknown Plant. From manual image interpretation, we know a crop of chickpeas was planted in May which grew until harvest in early September. Throughout the summer the model initially classifies this crop as wheat, soybean, or alfalfa. The post-processing correctly chose the Other Crop Type as the final class. In October and November there is confusion in the Dominant Cover category between soil and residue, even after post-processing. From the images we can conclude there was likely no activity in the field during this time, thus confusion likely stems from a moderate amount of residue on the field combined with low light conditions of this northern (47.7° latitude) site.

Crops going into dormancy in the winter and resuming growth in the spring are accounted for in the post-processing routines as demonstrated by the Konza Agricultural site in the NEON network in 2017 (NEON.D06.KONA.DP1.00042, Figure 6). A winter wheat crop (Figure 6B), which was planted in the fall of 2016, resumed growth in February. The remainder of that crop life cycle proceeded normally until harvest in July (Figure 6B). The crop type here is correctly classified as Unknown Plant by the classifier from January thru March here, since the plants were relatively small at this time. The correct classification of wheat began in March when the plants were large enough to confidently identify, and this was propagated back to the initial emergence in 2016. Additionally, the primary field at this site was harvested at the end of June 2017 (as seen in the original images), though the classification model indicated it happened mid-July (Figure 6A). This was due to the foreground plants being removed in mid-July, while the primary field was harvested in June.

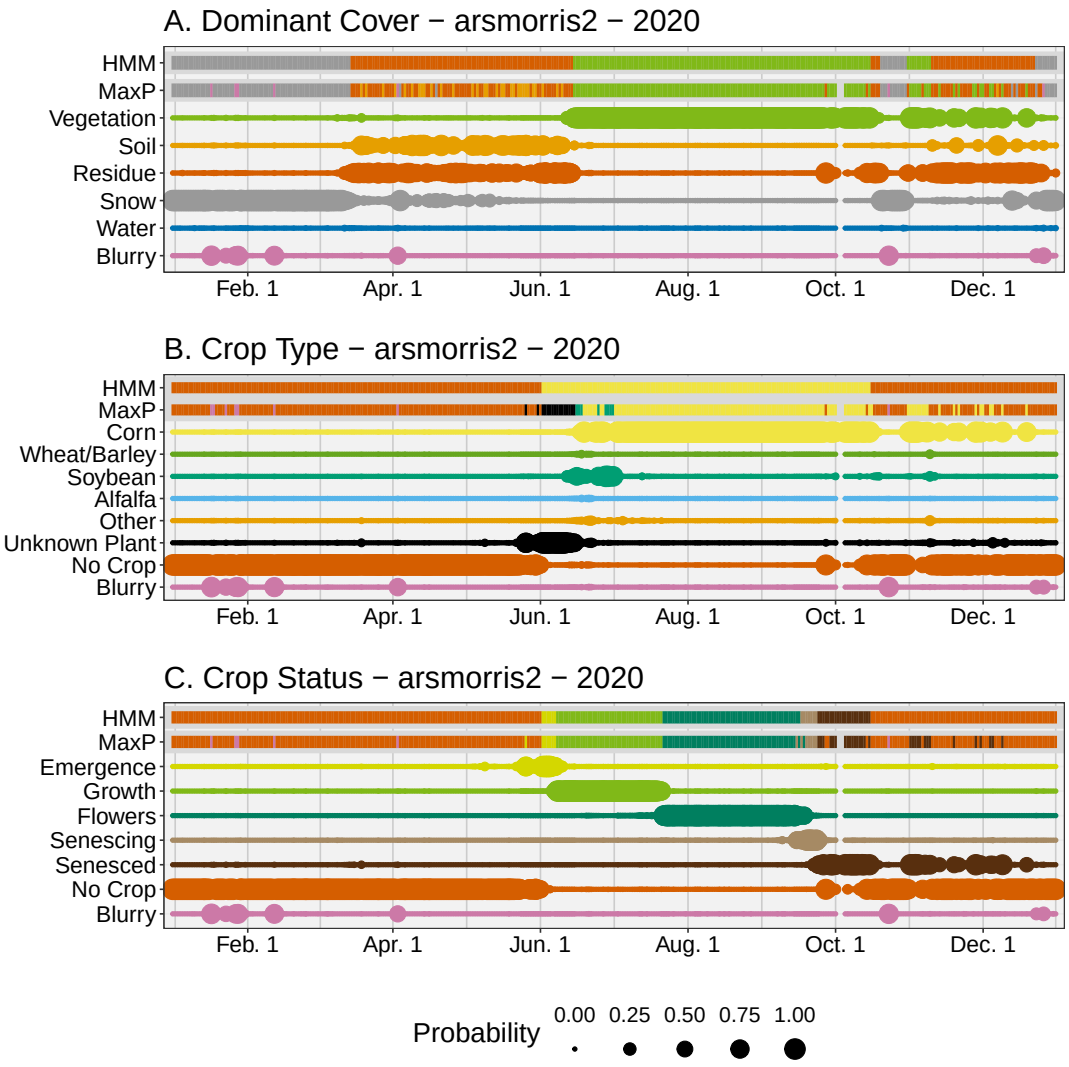

Figure 3: Classification results for the arsmorris2 site for the year 2020. The top two rows of each panel represent the final classification for either the daily maximum probability (MaxP) or the hidden markov model (HMM). The remaining rows in each panel represent the initial model classification for the respective class, where larger sizes represent higher probability.

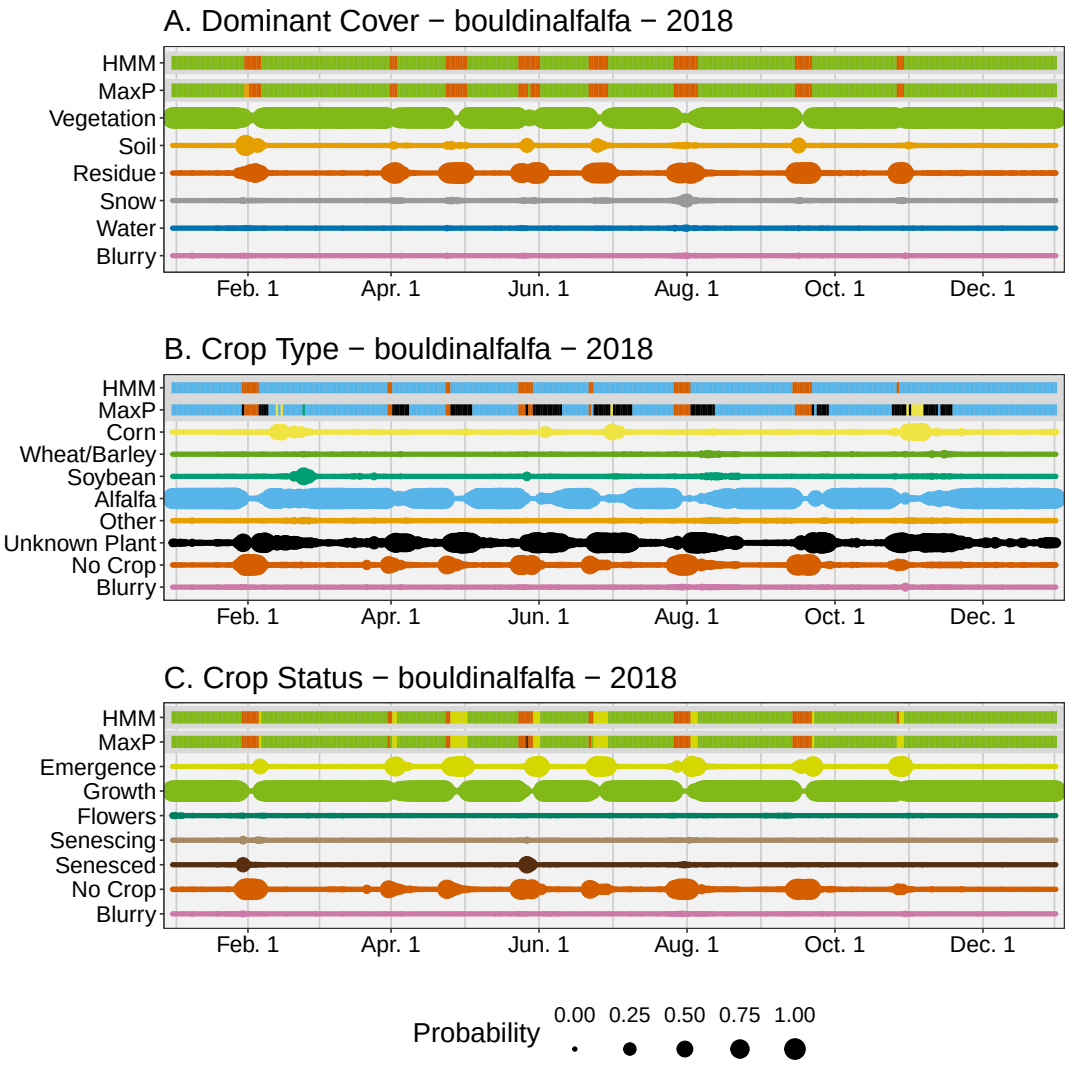

Figure 4: Classification results for the bouldinalfalfa site for the year 2018. The top two rows of each panel represent the final classification for either the daily maximum probability (MaxP) or the hidden markov model (HMM). The remaining rows in each panel represent the initial model classification for the respective class, where larger sizes represent higher probability.

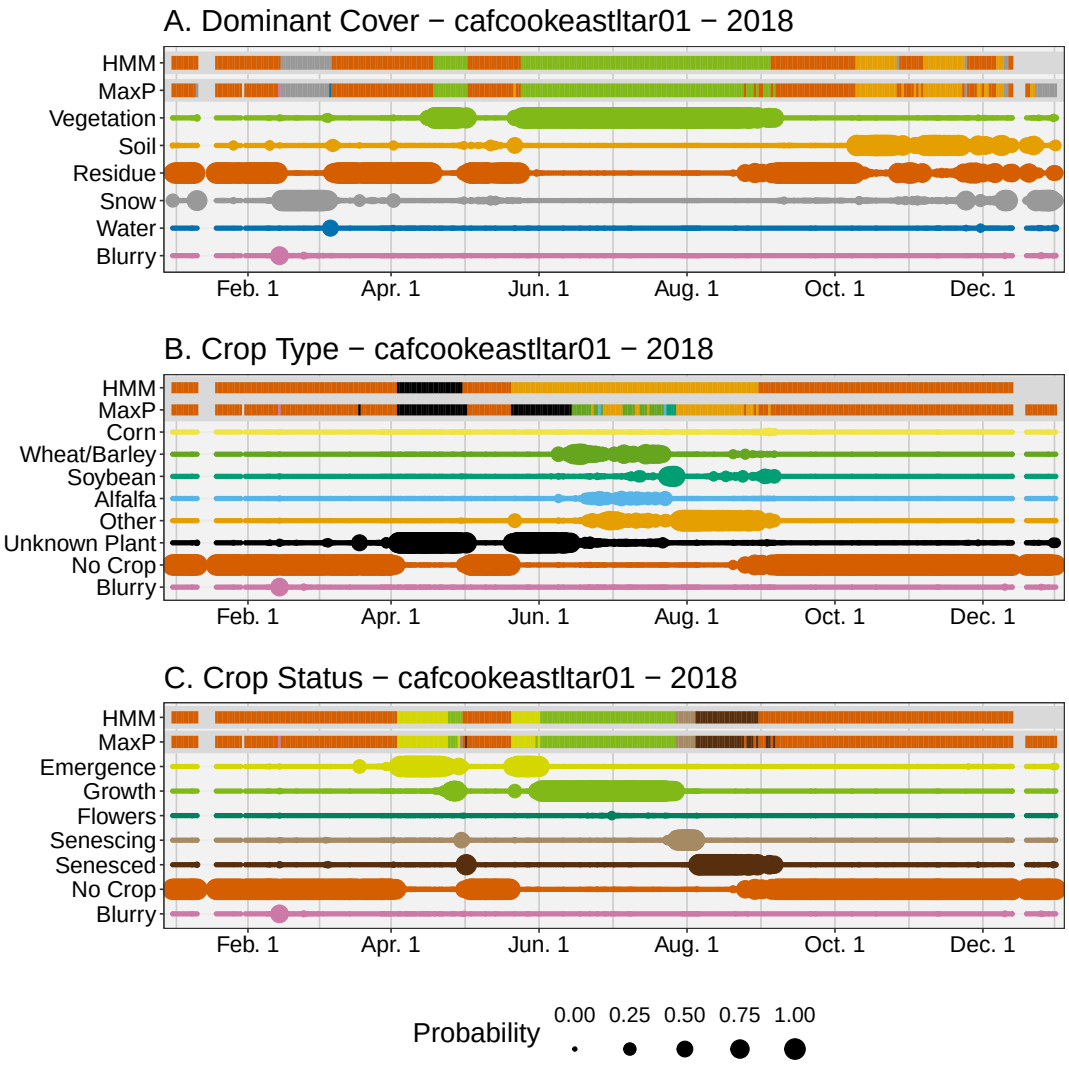

Figure 5: Classification results for the cafcookeastltar01 site for the year 2018. The top two rows of each panel represent the final classification for either the daily maximum probability (MaxP) or the hidden markov model (HMM). The remaining rows in each panel represent the initial model classification for the respective class, where larger sizes represent higher probability.

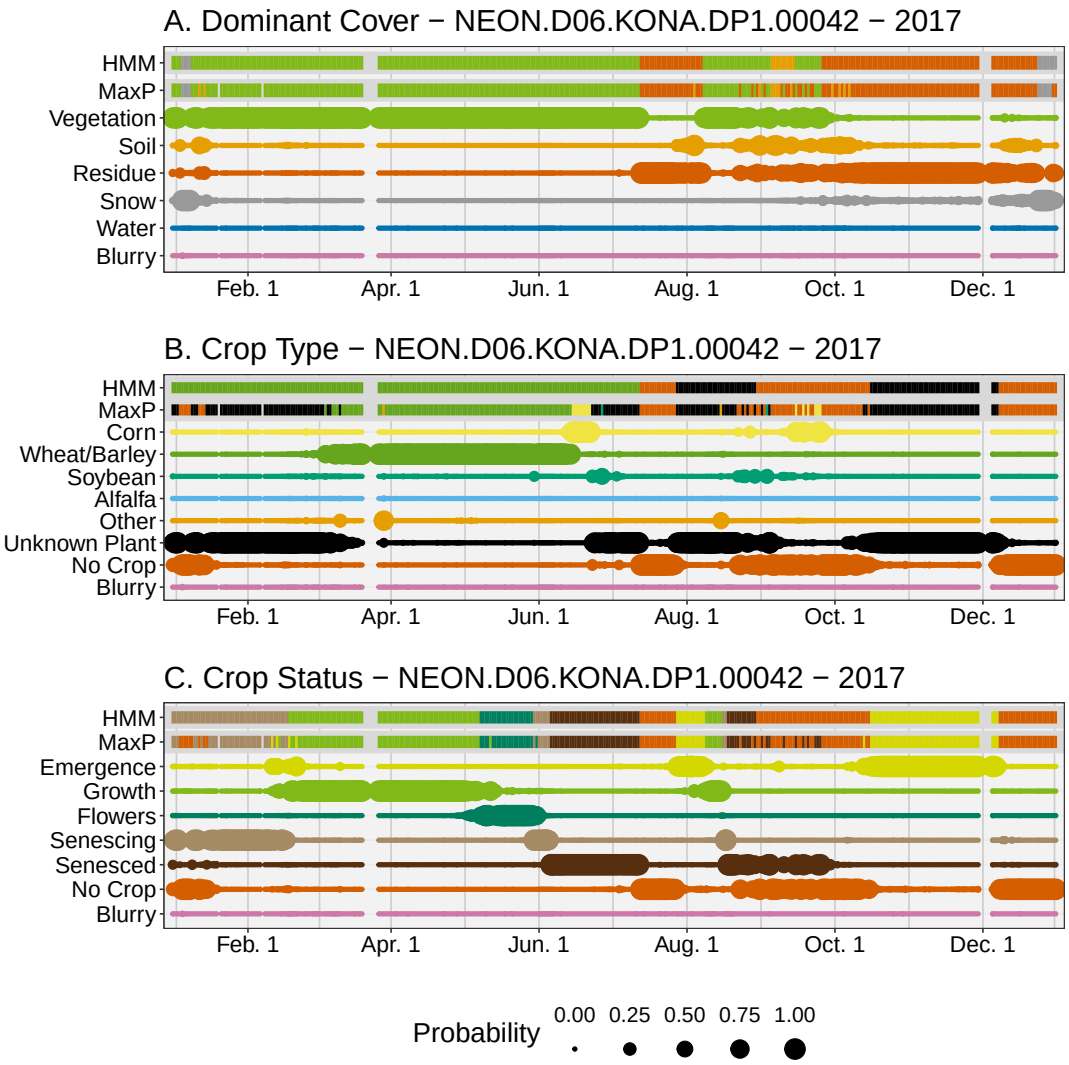

Figure 6: Classification results for the NEON.D06.KONA.DP1.00042 site for the year 2017. The top two rows of each panel represent the final classification for either the daily maximum probability (MaxP) or the hidden markov model (HMM). The remaining rows in each panel represent the initial model classification for the respective class, where larger sizes represent higher probability.

4 Discussion

Daily images from the PhenoCam network contain a wealth of information beyond just vegetation greenness, and here we showed they are also a novel source of cropland phenological information. Using a deep learning-based image classification model we identified the daily field state, crop type and phenological state from PhenoCam images in agricultural fields. Since mainstream classification models do not have a temporal component we applied a hidden markov model as a post-classification smoothing method which accounts for temporal correlation. This improved classification metrics and removed improbable transitions. Improvements would be beneficial to better classify field and crop states outside the primary growing season, and to better account for crops which go through a period of dormancy.

The classification model here was developed to simultaneously identify several crop and field attributes and has a variety of potential uses. For example the United States Department of Agriculture currently monitors crop status throughout the USA using surveys (*USDA-NASS*, 2012). An array of PhenoCams positioned in representative fields could enable a real-time crop status monitoring system using the methodologies outlined here. Remote sensing models for monitoring crop progression would benefit from the large temporal and spatial extent of PhenoCams in agricultural fields as a source of verification data (*Gao and Zhang*, 2021). The on-the-ground daily crop status data could also be used to parameterize or validate earth system models, where crop phenology is a primary source of error for crop yields (*Lombardozzi et al.*, 2020).

Numerous studies have used *Gcc* from PhenoCams to study various biological processes (*Richardson et al.*, 2018a; *Richardson*, 2019). Yet compressing images down to a single greenness index discards large amounts of information, especially in agricultural fields which are constantly managed (*Browning et al.*, 2021). Our approach here allows us to extract more relevant data from images, such as the crop type or the state of the field after vegetation is removed. These image-based metrics of crop type and stage from PhenoCam time series can complement *Gcc* as opposed to replacing it though. For example, *Gcc*, and other greenness metrics, can be used to derive the date of peak greenness or the rate of greenup or greendown, which reflects important plant properties not present in our image classification approach such as water stress and plant vigor (*Sakamoto et al.*, 2010; *Aasen et al.*, 2020). Added insight from complementary metrics enrich interpretation and offer decision-makers flexibility in crop management (*Browning et al.*, 2021).

Most studies using deep learning methods to identify cropland attributes use satellite or aerial imagery (*Benos et al.*, 2021), though several studies have used near-surface imagery similar to the work here. Yalcin 2017 used a CNN to classify crop types and had F1 scores of 0.74-0.87. Han et al. 2021 used a CNN to classify development stages in rice with F1 scores ranging from 0.25-1.0. The high accuracy seen here and in other studies shows the capability of tracking crop and field attributes with near-surface cameras. This approach is advantageous since the cameras are not affected by cloud cover and, after initial installation, do not have significant labor costs.

We identified several areas of our approach which could be improved. Firstly, the VGG16 model used here could be replaced with either a more advanced or a customized neural network model. Though the initial accuracy of the VGG16 model was relatively high, it was originally designed for

classification of common objects as opposed to croplands. It could likely be improved through model customization or fine-tuning of parameter estimation. Improving the initial image classification would improve the final results without any other adjustments to the post-processing routines.

Our approach here worked best during continuous periods where crops were present on the field. Once crops were removed, the dominant cover state could be difficult to discern due to soil, plant residue, and fully senesced plants having similar visual characteristics (Figures S2-S9). Improvements could potentially be made here by using a zoomed-in or cropped photo of the field. Since the images were compressed from their full resolution to 224x224 pixels, it is likely important details were lost. Han et al. 2021 showed that zoomed-in images, used simultaneously with full resolution images in a custom neural network model, greatly improved accuracy of rice phenology classification. Using zoomed-in images may also help with identifying the reproductive structures of crops. Though this may be limited by camera placement since even during manual annotation we could only identify reproductive structures of corn, wheat, and barley. Residue versus Soil classification may also be improved by classifying the amount of residue (e.g., the fractional cover of plant residue or soil) as opposed to using two distinct classes.

Our use of an HMM is an ideal solution to account for temporal correlation in the classified image time series. The progression of crops at a daily time step is constrained by plant biology, and these constraints are easily built into the HMM using the transition matrix. Additionally, the predicted probabilities from the classification model can be used directly in the HMM observation model, resulting in a straightforward data workflow. Since we used a basic HMM we had to create separate models for the Dominant Cover and Crop Status categories, which resulted in occasional inconsistencies. For example the Dominant Cover HMM may occasionally identify a time period as being predominantly vegetation, while the Crop Status HMM identifies the same time period as having no crop present (Figure 6). A multi-level, or layered, HMM may be able to overcome this by modelling the joint probabilities of classes across the two categories (*Fine et al.*, 1998). Temporal segmentation, a newer deep learning approach which is under active development, could model the joint probability of classes across the different categories in addition to having better performance than seen here (*Lea et al.*, 2016). A downside to temporal segmentation is that it would require fully annotated training sequences (i.e., annotations for all images in a year for numerous sites) as opposed to the random selection of training images used here.

We observed some mis-classifications when field management activities are not uniformly applied to all parts of a field in the camera field of view. In the NEON-KONA example (Figure 6) the final classification showed vegetation present due to foreground plants remaining even though the primary field was harvested. This could be improved by having a pre-processing step which identifies distinct agricultural fields within the camera field of view. Each agricultural field could then be classified independently. This would also allow the inclusion of PhenoCam sites focused on one to several experimental plots, which were excluded from this study. This step could be done automatically through image segmentation models, or manually as in the region of interest (ROI) identification in the current PhenoCam Network data workflow (*Richardson et al.*, 2018b).

Instead of discarding blurry or obscured images we accounted for them directly in the modelling. This is ideal since real-time applications must account for such images without human intervention.

The Blurry class across all three categories had high performance metrics for the training images, but with validation images it had the lowest performance among all classes. There are two possibilities for this low performance of this class. One is that the classification model was confident in classifying some partially obscured images as non-blurry where the human annotator was not (Figures S10-S11). Second was the low sample size of the blurry image class, which had less than 70 total mid-day images. This likely resulted in the over-fitting of the blurry class on the training images and resulting low performance among validation images. Obtaining more PhenoCam images which are blurry or where the field of view was obscured in some way would be beneficial, and could be obtained from the numerous non-agricultural sites. Regardless, the low accuracy of blurry images had little effect on the final results, since the final classification of any single day is determined by the joint classifications of all surrounding days in the post-processing.

Monitoring and assessing crop extent and status using a consistent, data-driven approach is essential to meeting the growing demand for food while meeting our sustainability goals in light of climate change. We formulated a workflow using a deep learning model applied to PhenoCam time series to generate a daily crop phenology time series for locations across the continental U.S. The workflow uses an HMM to account for the temporal correlation of daily images. The resulting outputs offer a ground truth to calibrate and refine existing models for mapping crop status and yield using satellite remote sensing.

Acknowledgments

We thank Andrew Richardson and Koen Hufkens for feedback on our methodology. This research was a contribution from the Long-Term Agroecosystem Research (LTAR) network. LTAR is supported by the United States Department of Agriculture. DMB was supported by CRIS 3050-11210-009-00D. The authors acknowledge the USDA Agricultural Research Service (ARS) Big Data Initiative and SCINet high performance computing resources (https://scinet.usda.gov) and funding from the Scientific Computing Initiative (SCINet) Postdoctoral Fellow program to support SDT. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. USDA is an equal opportunity provider and employer.

We thank our many collaborators, including site PIs and technicians, for their efforts in support of PhenoCam. The development of PhenoCam Network has been funded by the Northeastern States Research Cooperative, NSF's Macrosystems Biology program (awards EF-1065029 and EF-1702697), and DOE's Regional and Global Climate Modeling program (award DE-SC0016011). AmeriFlux is sponsored by the U.S. Department of Energy's Office of Science.

This research was supported in part by an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA). ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of USDA, DOE, or ORAU/ORISE.

References

Aasen, H., N. Kirchgessner, A. Walter, and F. Liebisch, Phenocams for field phenotyping: Using very high temporal resolution digital repeated photography to investigate interactions of growth, phenology, and harvest traits, *Frontiers in Plant Science*, *11*, doi:10.3389/fpls.2020.00593, 2020.

Abadi, M., et al., Tensorflow: Large-scale machine learning on heterogeneous systems, *arXiv*, 2016.

Barve, V. V., et al., Methods for broad-scale plant phenology assessments using citizen scientists' photographs, *Applications in Plant Sciences*, *8*(1), 754,275, doi:10.1002/aps3.11315, 2020.

Benos, L., A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, Machine learning in agriculture: A comprehensive updated review, *Sensors*, *21*(11), 3758, doi:10.3390/s21113758, 2021.

Browning, D. M., et al., Monitoring agroecosystem productivity and phenology at a national scale: A metric assessment framework, *Ecological Indicators*, *131*, 108,147, doi:10.1016/j.ecolind.2021.108147, 2021.

Chollet, F., Keras: The python deep learning library, *Astrophysics Source Code Library*, pp. ascl—-1806, 2018.

Conway, A. M., I. N. Durbach, A. McInnes, and R. N. Harris, Frame-by-frame annotation of video recordings using deep neural networks, *Ecosphere*, *12*(3), doi:10.1002/ecs2.3384, 2021.

Correia, D. L., W. Bouachir, D. Gervais, D. Pureswaran, D. D. Kneeshaw, and L. De Grandpre, Leveraging artificial intelligence for large-scale plant phenology studies from noisy time-lapse images, *IEEE Access*, *8*, 13,151–13,160, doi:10.1109/ACCESS.2020.2965462, 2020.

Esmael, B., A. Arnaout, R. K. Fruhwirth, and G. Thonhauser, Improving time series classification using hidden markov models, in *2012 12th International Conference on Hybrid Intelligent Systems (HIS)*, pp. 502–507, IEEE, doi:10.1109/HIS.2012.6421385, 2012.

Fine, S., Y. Singer, and N. Tishby, The hierarchical hidden markov model: Analysis and applications, *Machine learning*, *32*(1), 41–62, doi:10.1023/A:1007469218079, 1998.

Gao, F., and X. Zhang, Mapping crop phenology in near real-time using satellite remote sensing: Challenges and opportunities, *Journal of Remote Sensing*, *2021*, 1–14, doi:10.34133/2021/8379391, 2021.

Gao, F., M. C. Anderson, X. Zhang, Z. Yang, J. G. Alfieri, W. P. Kustas, R. Mueller, D. M. Johnson, and J. H. Prueger, Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery, *Remote Sensing of Environment*, *188*, 9–25, doi:10.1016/j.rse.2016.11.004, 2017.

Ghosal, S., D. Blystone, A. K. Singh, B. Ganapathysubramanian, A. Singh, and S. Sarkar, An explainable deep machine vision framework for plant stress phenotyping, *Proceedings of the National Academy of Sciences*, *115*(18), 4613–4618, doi:10.1073/pnas.1716999115, 2018.

Han, J., L. Shi, Q. Yang, K. Huang, Y. Zha, and J. Yu, Real-time detection of rice phenology through convolutional neural network using handheld camera images, *Precision Agriculture*, *22*(1), 154–178, doi:10.1007/s11119-020-09734-2, 2021.

Harris, C. R., et al., Array programming with numpy, *Nature*, *585*(7825), 357–362, doi:10.1038/s41586-020-2649-2, 2020.

Hufkens, K., E. K. Melaas, M. L. Mann, T. Foster, F. Ceballos, M. Robles, and B. Kramer, Monitoring crop phenology using a smartphone based near-surface remote sensing approach, *Agricultural and Forest Meteorology*, *265*, 327–337, doi:10.1016/j.agrformet.2018.11.002, 2019.

Jones, H. G., What plant is that? tests of automated image recognition apps for plant identification on plants from the british flora, *AoB PLANTS*, *12*(6), doi:10.1093/aobpla/plaa052, 2020.

Kosmala, M., A. Crall, R. Cheng, K. Hufkens, S. Henderson, and A. Richardson, Season spotter: Using citizen science to validate and scale plant phenology from near-surface remote sensing, *Remote Sensing*, *8*(9), 726, doi:10.3390/rs8090726, 2016.

Lea, C., A. Reiter, R. Vidal, and G. D. Hager, Segmental Spatiotemporal CNNs for Fine-Grained Action Segmentation, in *Computer Vision – ECCV 2016*, vol. 9907, edited by L. B., M. J., S. N., and W. M., pp. 36–52, Springer, Cham, doi:10.1007/978-3-319-46487-9_3, 2016.

Lombardozzi, D. L., Y. Lu, P. J. Lawrence, D. M. Lawrence, S. Swenson, K. W. Oleson, W. R. Wieder, and E. A. Ainsworth, Simulating agriculture in the community land model version 5, *Journal of Geophysical Research: Biogeosciences*, pp. 0–3, doi:10.1029/2019JG005529, 2020.

McKinney, W., Data structures for statistical computing in python, in *Proceedings of the 9th Python in Science Conference*, pp. 51–56, SciPy, Austin, Texas, USA, 2010.

Meier, U., *Growth stages of mono-and dicotyledonous plants*, Blackwell Wissenschafts-Verlag, 1997.

Milliman, T., et al., Phenocam dataset v2.0: Digital camera imagery from the phenocam network, 2000-2018, doi:10.3334/ORNLDAAC/1689, 2019.

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning, *Proceedings of the National Academy of Sciences of the United States of America*, *115*(25), E5716–E5725, doi:10.1073/pnas.1719367115, 2018.

Python Software Foundation, Python language reference manual, version 3.6, url-http://www.python.org, 2003.

R Core Team, R: a language and environment for statistical computing, 2017.

Richardson, A. D., Tracking seasonal rhythms of plants in diverse ecosystems with digital camera imagery, *New Phytologist*, *222*(4), 1742–1750, doi:10.1111/nph.15591, 2019.

Richardson, A. D., K. Hufkens, T. Milliman, and S. Frolking, Intercomparison of phenological transition dates derived from the phenocam dataset v1.0 and modis satellite remote sensing, *Scientific Reports*, *8*(1), 5679, doi:10.1038/s41598-018-23804-6, 2018a.

Richardson, A. D., et al., Tracking vegetation phenology across diverse north american biomes using phenocam imagery, *Scientific Data*, *5*(1), 180,028, doi:10.1038/sdata.2018.28, 2018b.

Sakamoto, T., B. D. Wardlow, A. A. Gitelson, S. B. Verma, A. E. Suyker, and T. J. Arkebauer, A Two-Step Filtering approach for detecting maize and soybean phenology with time-series MODIS data, *Remote Sensing of Environment*, *114*(10), 2146–2159, doi:10.1016/j.rse.2010.04.019, 2010.

Schreiber, J., Pomegranate: fast and flexible probabilistic modeling in python, *Journal of Machine Learning Research*, *18*(164), 1–6, 2018.

Seyednasrollah, B., A. M. Young, K. Hufkens, T. Milliman, M. A. Friedl, S. Frolking, and A. D. Richardson, Tracking vegetation phenology across diverse biomes using version 2.0 of the phenocam dataset, *Scientific data*, *6*(1), 222, doi:10.1038/s41597-019-0229-9, 2019.

Simonyan, K., and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv*, 2014.

USDA-NASS, The yield forecasting program of NASS. Report SMB 12-01, *Tech. rep.*, USDA National Agricultural Statistics Service, 2012.

Wehmann, A., and D. Liu, A spatial–temporal contextual markovian kernel method for multitemporal land cover mapping, *ISPRS Journal of Photogrammetry and Remote Sensing*, *107*, 77–89, doi:10.1016/j.isprsjprs.2015.04.009, 2015.

Weinstein, B. G., A computer vision for animal ecology, *Journal of Animal Ecology*, *87*(3), 533–545, doi:10.1111/1365-2656.12780, 2018.

Weiss, M., F. Jacob, and G. Duveiller, Remote sensing for agricultural applications: A meta-review, *Remote Sensing of Environment*, *236*, 111,402, doi:10.1016/j.rse.2019.111402, 2020.

Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, 2016.

Wickham, H., et al., Welcome to the tidyverse, *Journal of Open Source Software*, *4*(43), 1686, doi:10.21105/joss.01686, 2019.

Yalcin, H., Plant phenology recognition using deep learning: Deep-pheno, *2017 6th International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2017*, doi:10.1109/Agro-Geoinformatics.2017.8046996, 2017.

Zeileis, A., and G. Grothendieck, zoo : S3 infrastructure for regular and irregular time series, *Journal of Statistical Software*, *14*(6), doi:10.18637/jss.v014.i06, 2005.